CW02 - A Tiny Data Science Project

*A brief **overview of the domain** of the data being analysed. Describes **problems** to be tackled and clearly list any **analytical questions** that are tackled. For example, try to formulate questions (or abstract tasks) such as, "We like to see differences between group X and group Y". In light of these tasks, list your **objectives** in the analysis and briefly describe your **analysis strategy**. Do not forget to clearly reference to the resources where you get your data from. (**Maximum 2 pages**).*

Book-Crossing Dataset (BCD) has been collected by Cia-Nicholas Zieglar, at DBIS Freiburg, during 4 weeks work. The data was collected using a web crawling algorithm on the Book-Crossing online community. The datasets have been acquired directly from the DBIS Freiburg University website, thanks to Cai-Nicolas Ziegler et al (2005) for the data. The BCD dataset comprises of three structured CSV files; in summary there are over 1.1 million book ratings by over 270,000 unique users. Each file contains a few individual dimensions. The three files; BX-Users, BX-Books and BX-Book-Ratings, contain only anonymous data for each of the reviews where no name is stored, only a unique user ID. A book review can typically thought of as an executive summary review of a book, usually in the form of free text or a numerical rating index. The BCD does not contain the free text feedback, the rating are stored as values, in a structured dataset. The values are represented on a 0-10 scale, where 10 is the best result.

The objective of the analysis is to understand if book reviews can be utilised to understand behavioural or demographic similarities between users. The list below contains a series of questions about the data that would like to be understood:

- Can book reviews be used to understand feedback habits of the reviewers in order to enhance recommendation for all users? This knowledge could also be used by Publishers to understand if a book has been successful or not.
- Does the length of a book title effect the distribution of book reviews? This is an exploratory analysis to see if the length of a title is statistically significant in the average review.
- Does the year of publication imply popularity? If any behavioural groups exist, do they tend to review certain books?
- Is the spread of user types even across each country? This will help both BC and Publishers to anticipate what types of reviews will be created by each market, therefore gain better insight of the successfulness of a book.
- Finally do particular Publishers attract certain types of book reviewers?
- Finally for further analysis, if time permits, can a Recommendation System be introduced to utilise the analysis above as a means of fine-tuning the system. The intention will be to use either a Support Vector Machine (SVM) or a pre-compiled Python Recommendation algorithm, like the python-recsys, for the support of this. Alternatively is it possible to develop a predictive model that is capable of determining if a reviewer will be a within one of the classes.

The analytical approach, through the use of a clustering technique, will be tried first to establish a means of classifying the book reviewers into particular groups (K-Means, SVM) to see if any groups can be found in the data. Prior to identifying groups the data will be cleaned and combined into one dataset, such that credible data is retained

and un-realistic data is excluded from the analysis. An example of this may be that the user has entered an age over 100 and that they live on Mars. Once the pre-processing has been obtained the intention will be to systematically try and derive analysis based on the analytical questions described above, checking at each stage for statistical significance and making note of key insights if any have been obtained.

For the analysis a range of machine learning techniques to classify any groups has been developed and the intention is that once the analysis has been compiled the use of visualisations will be produced to visually represent if the analysis is successful or not.

*Document any steps taken to get the **data ready for analysis**, detailing on steps such as handling missing data, merging and transforming data, etc. (**Maximum 1 page**)*

The data wrangling was primarily conducted in Python using Pandas and Numpy for manipulations and mathematical operations. The data contained over 278,000 unique reviewers' details, however the after filtering out the un-realistic data the remaining data consisted of over 162,000 reviewers. The steps to filter out the non-credible items of data have been outlined below.

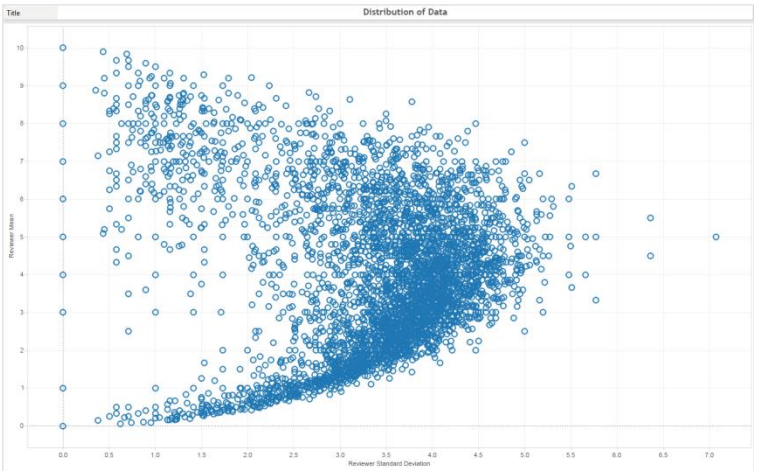
The user dataset contained minimal attributes about the users; only their age, location and a unique identifier. In order to find credible reviewers, the initial stage was to remove users that had not entered their location accurately. The first step was to find if the location field had been completed correctly. A correctly populated location would be one completed with a town, state and a country. This was verified by counting the number of commas in the field; if the count was two the field may have been completed correctly. At this stage filtering out countries that are not real was not conducted. The second step was to filter out ages that were un-realistic, I made the decision that reviewers would be in most likely be in the age range of 15 to 80, while there may be users that could be just outside of this range they have been discredited as I made the decision that these could be phony and that could lead to biased results.

Once the user data had been cleansed for non-credible items, the next stage was to merge each of the datasets into one main Data-frame. I made use of left joins which meant that the use of dropping null values could be utilised to remove the pre-filtered work mentioned earlier. The final stage of data preparation was to obtain a list of the countries, correctly formatted and proper case, that if required for visualisation later would easily be interpreted by tableau. This data included the Longitude and Latitude of the Countries, giving an approximation of the user's location, meaning that geographical plots could be added later on. The consequence of processing the data this way meant that null values would be discounted as well as any mock values, Mars or Venus for example.

This concluded the pre-processing of the data however at nearly every stage of the analysis, transformations are used to add derivations to the master data-frame (merging), calculation summary statistics and also the creation subsets of the dataset for detailed analysis on a subset of the data.

*Document the steps taken in your analysis, provide details on the analytical tools used and explain your choices. Try to fulfil most of the analytical process outlined and follow the suggestions listed above. Make use of visualisations where appropriate. (**Maximum 3 pages**)*

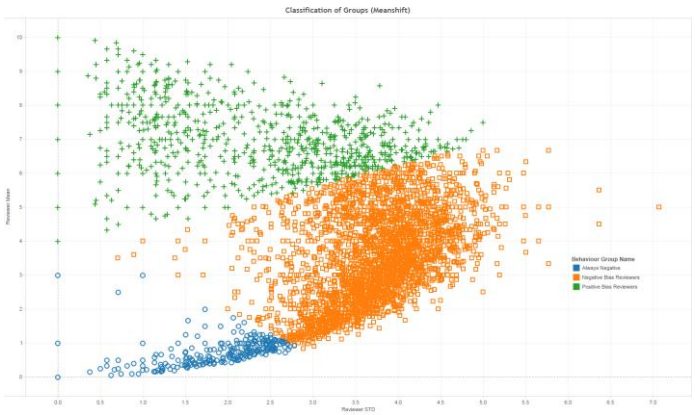
Once the data had been prepared, the first step was to see if there were any identifiable groups based on the user's average score and standard deviation. It was decided that these two variables would define the characteristic behaviour of someone reviewing a book. It was hypothesised that individuals could be grouped into at least 2 different groups. The first group could be those that only review books that they feel were worthwhile reading and the second group was those individuals that only reviewed with negatively. Plotting on a scatter graph meant the distribution obviously exhibited the potential have this characteristic.



From visually analysing the graphic it could be actually said that the data splits into at least 3 groups. Further analysis was undertaken to see if there was any mathematical similarity that could mean these groups could be identified.

The methods experimented with were the K-Means and Meanshift, both feature space analysis techniques that can identify shapes within the data. The reason why these methods were chosen was to employ robustness in the findings. From the Meanshift plot below it can clearly be seen that there are 3 distinct groups, however when experimenting with K-Means I identified 3-5 possible groups in the data. It was decided that the three groups identified using the Meanshift method could be labelled as; 'Positive Bias Reviewers', 'Always Negative' and 'Negative Bias Reviewers'.

Meanshift was chosen over K-Means as there was no drawback on having to previously define the number of clusters. This stage of the process involved deriving new variables from the original datasets, namely the user average and standard deviation.



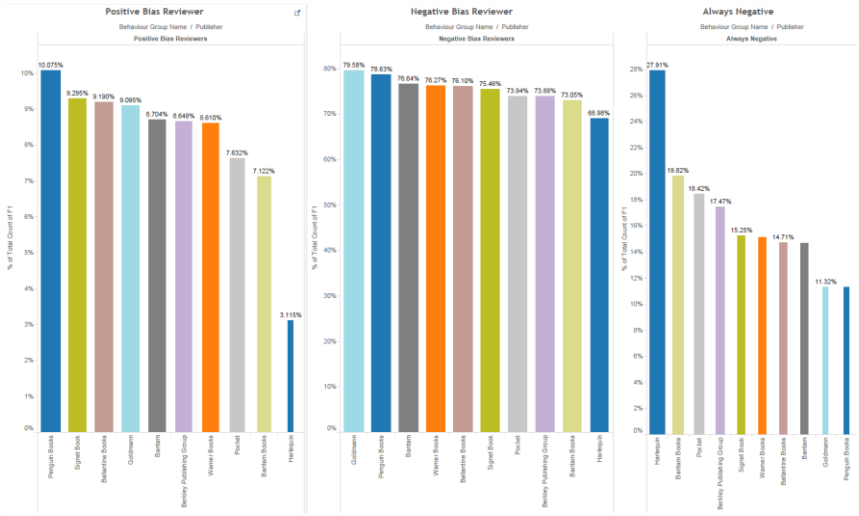
It is important to note that if no groups were found then the objectives of the analysis would have had to be modified or corrected as a result. As discussed above, 3 distinguishable groups have been identified from the plot above, however with further analysis more groups could potentially be identified making this a very interesting group of people with a standard deviation of greater than 4. For the purpose of further analysis this was not taken into consideration, but it is certainly one thing that would be of interest if time permitted. The classifications of the three groups that have been described above are what has been used throughout the remaining analysis.

The first objective that was conducted was to see if there were any differences between the groups and the varying lengths of a book title, a visual analysis can be seen in the Tableau graphic below. Also confirmed using a One-Way ANOVA, was a statistical difference between groups. This stage involved the creation of new variables, a function to convert the code to a label and also merging. This analysis could also be used to identify sub-groups of the three derived above. For instance, if I am regularly reviewing books with long titles (could be Journals or Technical Texts) am I potentially a more reliable reviewer that that of someone who reviews a long titled text as a one-off occurrence.

The second objective was to understand if there were any differences between the groups when considering if the age of the book affects its rating. From the graphic below there appears to be no significance in the age as and no skew towards newer or older books. The graphic further begs questions about cohorts of users that may share similarities between each other, this could lead to potential gain for both retailers and publishers established what is trending and what types of users they are attracting to their site or books.

The third and fourth objectives were to understand if there are any differences between ratings, countries and publishers. The visualisations are shown below:

The size of the pie indicates the number of users within the group



The final objective was to see if the analysis could be used conjunction with a recommendation system which could be useful for an end-user and potentially for a more responsive system. This was not deemed relevant and proved difficult to test during the work, the pertinent reasons include the lack of customers to test on and the focus of the analysis had been on the service providers and not the users. This final objective was modified to enable the identification of the 'Always Negative' reviewers, the benefit of this predictive capability is discussed later but even from the limited quantity of data this was found to be very accurate when evaluating the method over multiple validation sets. With regards to the techniques utilised; derivation of new variables, conversion of categorical labels to integers and the logistic functionality from a machine learning module were all utilised in this analysis.

The model that has been used is shown below above as well as the output results:

Model:

Always_Negative_Boolean ~ Book_Rating + Age + Reviewer_Mean + Reviewer_STD + Reviewers_N + Country + State

Results:

The Mean Accuracy for the 10 Folds:	99.39%
The Mean F1 for the 10 Folds:	97.84%
The likelihood of a User Being in the "Always Negative" user class:	14.34%

The results obtained above are highly accurate with regards to being able to identify the three types of users; this type of capability can be used recursively to update the results each time a user returns and reviews another book or in identifying the group of a new user. Deriving meaningful results is not easy using this method; a more appropriate method might be the Random Forests or Decision Tree Analysis to find the characteristics that make up those in the Always Negative group. Operationally in a company, the size of dataset is likely to increase daily; therefore scaling and over fitting would be an issue as it is more intensive to tuning to those methods in comparison to the Logistic Regression method.

*List your findings briefly, reflect on to what degree the objectives have been met. Briefly speculate on how your findings can be turned into relevant business and/or scientific value. Make use of visualisations where appropriate. (**Maximum 1 page**)*

The analysis show has been conclusive to the findings that there are different groups and the groups have been defined by the similar characteristics of users reviewing the data. It has also been shown that a predictive Logistic Regression model can be utilised as a binary classifier to determine if a review would be within that class or not. The purpose of using this method over the clustering technique has been discussed, but in summary it is computationally more difficult to compute the K-Means and Meanshift when the volume of data increases. Additionally it is also possible to tune/update the LR model as new data is created and the profile of a user changes in time. With regards to the objectives of the project, these have been assessed, where required they have been adjusted to be within the capability and feasibility of myself and the testing resources available and results have been obtained.

Real-World Applications of this type of Analysis:

Audience Segmentation can be used for marketing and content analysis, it could be also be used in conjunction with text analysis to understand the sentiment and topic of the review. This type of classification of the groups could help both retailers and publishers (including Authors) understand what makes a book successful through detailed analysis of each of the reviews and groups to identify features of a book that make it successful or not. Retailers would likely only want to show a unbiased results, the reason for this is that they are financially driven and so having reviews on the site that have been provided by unbiased reviewers could help lower potential revenue on a book. Contrastingly, a Publisher would like to understand the whole spectrum of reviewers; this would help them to target specific audiences or help decide on which new authors work to publish.

Cohort Analysis can also be used with the type of classification methods mentioned in this analysis which would be very useful for companies to understand how the view of its customers evolve over time. Financially this is of most benefit to the retailers and publishers as it will allow them to fine tune their products to a specific customer audience. One of the most prolific uses of customer profiling is by Amazon [4] who are able to tailor their service for each customer to increase sales and profit.

*Describe the tools you've used. This could be either talking about your Python code and the packages you've used or any other software/system you've developed and/or utilized. (**Couple of paragraphs**)*

For the majority of this project I have solely used Python and primarily the following modules; Pandas, Numpy, Scipy, as well as many elements of the sklearn module (Pre-processing, Logistic Regression, Cross Validation, K-Means and Meanshift). The reason for this was the impressive array of modules available meant that the functionality has been pre-built which has made it possible for me to apply them effectively to the Book-Crossing Dataset. There is great documentation and examples online that have helped with debugging and learning as I went through each stage of the objectives.

Where possible I tried to make use of Tableau for visualisation of the data. Tableau offers the capability to manipulate visuals a lot quicker, in comparison to Matplotlib, to enable experimentation with different types of representations of the data. This would not be otherwise possible in Python given that it is all processed by the command line. Not only is the data visually more compelling in Tableau, it has also the capability of interactions. Adding the ability to drill into the data helps to identify further questions about the data as well as gain new insights. If more time was available, I would have liked the opportunity to experiment with the Nodebox [3] and Bokeh [2] packages for developing interactive visualisations similar to Tableau, as a pure end to end analysis in Python.

References:

[1] - Improving Recommendation Lists Through Topic Diversification,

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; *Proceedings of the 14th International World Wide Web Conference (WWW '05)*, May 10-14, 2005, Chiba, Japan. *To appear.*

[2] - Bokeh Module for Python

Continuum Analytics. (2014). *Welcome to Bokeh.* Available: http://bokeh.pydata.org/index.html. Last accessed 13th Dec 2014.

[3] - Nodebox Module for Python

Experimental Media Research Group. (2014). *Nodebox.* Available: https://www.nodebox.net/. Last accessed 13th Dec 2014.

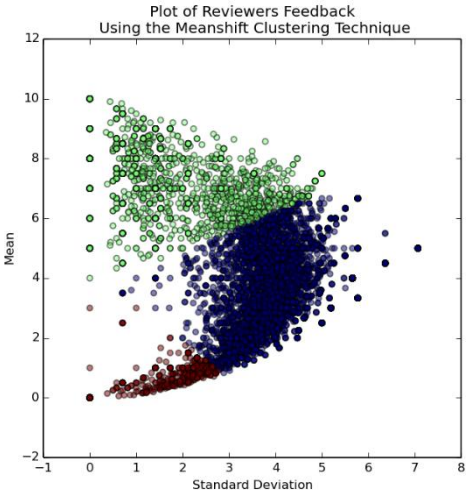
[4] – Ravi Kalakota. (2011). *Big Data Analytics Use Cases.* Available: http://practicalanalytics.wordpress.com/2011/12/12/big-data-analytics-use-cases/. Last accessed 13th Dec 2014.

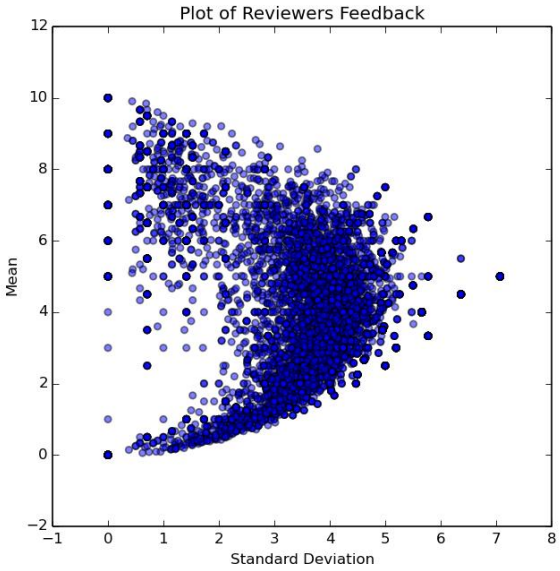
Appendix:

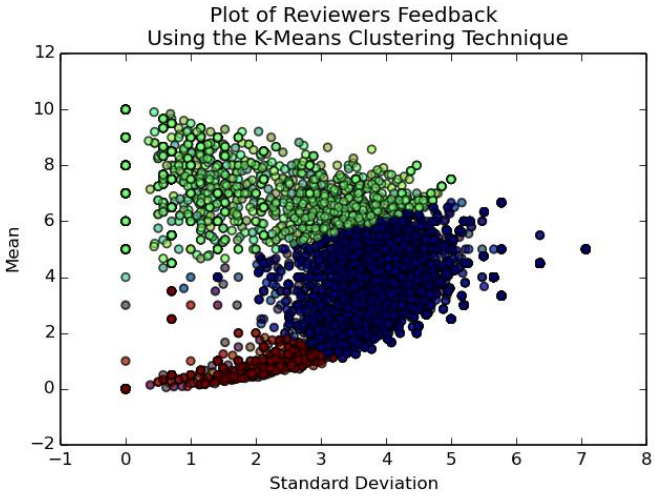
Python Code can be found at: https://github.com/dandxy89/CityUniversity2014/blob/master/IMN432-CW02

Book Crossing Data can be found at: http://grouplens.org/datasets/book-crossing/

Additional Graphics:







Useful Sites:

https://bitly.com/bundles/o_n4jfgiuuf/2 - Data Science Primer resources that has been put together by Zipfian Academy.